RECOGNIZING EMOTIONS OF CHARACTERS IN MOVIES

Ruchir Srivastava, Shuicheng Yan, Terence Sim

National University of Singapore,
Singapore

Sujoy Roy

Institute for Infocomm Research, A*STAR
Singapore

ABSTRACT

This work presents an investigation into recognizing emotions of people in near real life scenarios. Most existing studies on recognizing emotions of people have been conducted under controlled environments where the emotions are not spontaneous, rather highly exaggerated, and the number of modalities considered and their interactions is limited. The proposed bimodal approach fuses facial expression recognition (FER) with the "semantic orientation" of dialogs of actors to identify emotions under difficult illumination conditions, pose variations and occlusions in scenes. Experiments conducted on a dataset of 700 video clips from 17 movies demonstrate that the proposed fusion approach improves emotion recognition performance over unimodal approaches.

Index Terms— Bimodal Emotion Recognition, Semantic analysis, Fusion

1. INTRODUCTION

Face is the index of the mind. Combined with that, *what we speak* and *how we speak* or *react to external stimuli* are all helpful cues in making in-roads into understanding the complex world of human emotions. The ability to recognize emotions of people (ER) in real life has been identified as the essential component of many applications viz. social robotics, interactive recommendation systems etc. Most existing works on ER have analyzed emotions of people in controlled laboratory environments. Recent studies [1] indicate the need for developing ER approaches suitable for day-to-day environments to build useful products. The work proposed in this paper aims at recognizing emotions of people in near-real life scenarios, namely movies (Fig. 2(a-d)).

Why movies? While the best case scenario would be to capture data of people in day-to-day life, like using surveillance cameras, casual home video footages; such data can be extremely noisy and challenging to work with. For example, in most of such natural videos the face itself would be barely visible because no one in their day to day life would pose for the camera. Hence getting natural data to train such systems would be very hard to get. On the other hand movie data comes close to imitating day-to-day scenarios. The audio-visual quality can be extremely challenging in terms of occlusions, noisy environments, pose variations etc. However, compared to audio-visual data from day-to-day life, movie data has editorial advantages. For example, the camera mostly focuses on faces of actors in frontal or near frontal positions thus facilitating facial expression recognition, and a fair bit of associated information in the form of dialogues, script, clear audio, is available. Hence we refer to movie data as near real life data.

The available cues for ER of humans are mainly in the form of facial expressions (FER), body gestures, speech acoustics and lexical cues from dialogs (LC). Extracting relevant information from visual cues becomes challenging due to variations in image appearance like poor illumination, pose variations, occlusions and so forth, some of which are shown in Fig. 2. Speech acoustics are affected by background noise. Movie dialogues are short and succinct and this makes lexical analysis of dialogues difficult (as explained in section 3).

This work presents a dynamic weighting based multimodal approach (Fig. 1) to combining FER with LC for recognizing emotions of actors in movies. A novel method for identifying lexical cues is proposed that is particularly suited for movie dialogues. Emotions are classified as positive or negative. The two class prediction can be seen as a higher level abstraction which gives valuable insight into a more challenging problem when there are many more emotion categories.

2. STATE-OF-THE ART

Existing approaches on ER of people are not robust enough for difficult data sets like movie data. Several works on ER have used speech acoustics from audio as features[2]. However, information extracted from acoustic cues is noisy, not just due to contextual noise but also the huge variability in how people speak. Some works have reported that lexical cues perform better for ER as compared to acoustic cues [3].

Lexical features have been used in conjunction with visual cues in existing literature. Schuller et al. [4] have detected human interest level combining different cues including lexical features from speech. For linguistic analysis, a vocabulary of terms of interest has been established which includes non-linguistic vocalizations such as coughing, yawning, laughter, consent, hesitation and words such as *if*, *oh*, *yeah* and so on.

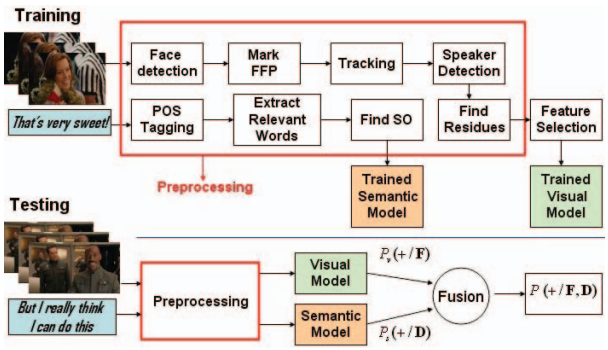

Fig. 1: Framework for the proposed system

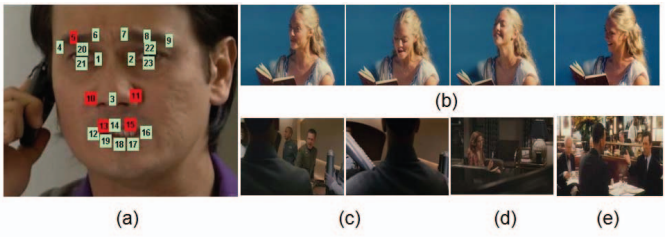

Fig. 2: a) 23 FFPs marked on the face. Five most discriminating points as found out by feature selection are shown in red. Few difficulties in using visual cues for ER in natural environments: b) Head motion c) Face Occluded in frames; d) and e) Small face size. The proposed approach tries to recognize emotions in such difficult scenarios.

In ER from dialogues, these terms may be insufficient. For example, for two phrases, *"Oh! Nice painting"* and *"Oh! It is very troublesome."*, occurrence of the word *'Oh!'* indicates interest of the speaker but emotion is conveyed by the words *'Nice'* and *'troublesome'*.

The concept of *emotional intelligence* has been used by Lee and Narayanan [5] for ER from dialogues. Selected words are analyzed separately which may not give sufficient clue about the context in which the word was spoken. Turney [6] has performed sentiment mining in movie reviews using word pairs. We note that techniques tested on movie reviews cannot be directly applied to recognizing emotions from dialogues (refer section 3).

Also note that the problem addressed in this work is not the same as detection of "affective nature of movie scenes"[7] and what emotions it arouses in viewers. We are interested in individual and collective emotions of actors in the movie which may or may not have predictable effect on viewers. For example, some action movie scenes make people laugh.

3. EMOTION RECOGNITION FRAMEWORK

We propose a bimodal framework for emotion recognition combining facial expression recognition with lexical analysis of dialogues in movies.

	Proposed method	Turney's Method [6]
Failures	15%	71.9%
Accuracy	86.7%	23.2%

Table 1: Comparison of proposed method and method in [6] for ER using Semantic Orientation (SO). Failure rates indicate failures in extracting emotional words. Accuracy refers to emotion prediction accuracy.

3.1. Facial Expression Recognition (FER)

Given video frames, faces are detected[8] in the first frame along with 23 facial feature points (FFPs). Next, FFPs are tracked in subsequent frames using a PPCA based algorithm [9]. Speaker detection is performed[10] and for a speaker, displacement of FFPs from neutral are used as features. Discriminative FFPs are selected using Fisher ratio test. A frame is labeled positive or negative using an SVM classifier learnt over examples of positive and negative classes of expressions. Majority voting is used to determine the label of the sequence.

3.2. Lexical Analysis of Dialogues

The lexical cues of dialogues are computed using an approach that determines their *semantic orientation* (SO). SO is indicative of the speaker's expression being positive or negative. Turney[6] proposed a rule-based algorithm for finding SO for "movie reviews". Movie reviews mostly contain sentences where grammatical structures are identifiable for which such rules apply. In comparison dialogues are very short and do not contain well-defined semantic structure, thus inhibiting such rules to be applied.

We propose a rule-based algorithm for determining the semantic orientation of movie dialogues. The key to computing the SO, lies in extracting emotional words or word pairs. From the dialog, word-pairs were extracted whose parts of speech followed any one of the following patterns: (1) Adjective-Noun (2) Adverb-Adjective (3) Adjective-Adjective (4) Noun-Adjective (5) Adverb-Verb (6) Verb-Noun. For example, in the dialog "*John wore a nice shirt and was looking handsome*"; the word pair "*nice shirt*" is an example of Adjective-Noun Pair and thus will be extracted as an emotional word pair. Once word pairs are extracted, from the remaining dialog, all the adjectives, adverbs and nouns (except proper nouns) are extracted as emotional words (The word *handsome*; being an adjective; is extracted as an emotional word). Each of the words extracted either alone or as word pair is fed to the PMI-IR algorithm[6] which calculates SO for that word. The words extracted in the example will be *handsome*, *nice* and *shirt*. Note that words in a pair are separated as individual words before feeding them to the PMI-IR algorithm. The basic idea behind the PMI-IR algorithm is that

a word with negative connotation is more likely to co-occur with a negatively connotated word rather than a positively connotated word. Co-occurrence for a query word or phrase is calculated using internet search engines. SO for the dialog is calculated as the mean of SO values for all the extracted words.

To determine the probability of a dialogue to be positive/negative, SO magnitudes were normalized by the most positive (SO_+) and most negative (SO_-) SO values. SO_+ and SO_- are found from the training data. Thus, the probability of a dialogue $\mathbb{D}$ with SO value S being positive is given by $P_s(+/\mathbb{D}) = S/SO_+$ if $S \geq 0$ and $P_s(+/\mathbb{D}) = 1 - S/SO_-$ if $S < 0$.

3.3. Fusing cues from FER and lexical analysis

The combined probability of a video frame having positive/negative emotion is computed by a "weighted" sum of probabilities obtained using visual and lexical cues. The probability of the sequence having positive/negative emotion is given by the mean of probability values over all the frames.

Finding Weights: Weights given to the individual cues are proportional to the confidence in that cue. To calculate weights for visual cues, factors were identified which can affect the performance of ER from visual cues. Factors and associated parameters for frame i of a video sequence are as follows:

Pose(p_1^i,p_2^i) p_1^i is the ratio of horizontal distances between the nose tip and each of the outer eyebrow corners. Greater of the two ratios is considered. For a frontal face $p_1^i \approx 1$ and any deviation from 1 indicates rotation of the face in yaw direction. p_2^i is the slope of line connecting the two outer eyebrow corners. $p_2^i \approx 0$ for a frontal face and deviation from zero shows rotation of the face in roll direction.

Scale(p_3^i) Size of the face (Section 4). This distance should be large for a good recognition performance.

Intensity of expression(p_4^i) Sum of the magnitudes of residues for the frame, normalized by p_3^i. Higher value of p_4^i improves recognition.

Head motion(p_5^i) This gives the displacement of the reference (nose tip) from its position in the first frame. Note that the nose tip is least affected by expressions. For a good recognition performance, p_5^i should be low.

Weight given to the visual cue for a frame i is calculated as

$$w_v^i = \frac{\frac{\alpha_v^i}{\alpha_{max}}}{(\frac{\alpha_v^i}{\alpha_{max}} + \alpha_s)} \quad (1)$$

$$\text{where } \alpha_v^i = \frac{p_3^i p_4^i}{(1+a_1^i|p_1^i - 1|)(1+a_2^i|p_2^i|)(1+a_3^i p_5^i)},$$

α_s is the normalized SO magnitude and α_{max} is the maximum value of α_v^i for the training data. The weight for lexical cues is given by $w_s^i = 1 - w_v^i$.

	P	N
P	73.3	26.7
N	20.0	80.0

(a) Visual (ARR=76.7%)

	P	N
P	86.7	13.3
N	13.3	86.7

(b) lexical (ARR=86.7%)

	P	N
P	90.7	9.3
N	5.3	94.7

(c) Fusion (ARR=92.7%)

Table 2: Effectiveness of fusing visual and lexical cues. Confusion matrices for classification a) Using visual features alone, b) Using lexical analysis alone, c) By fusion ARR: Average Recognition Rate; P: Positive; N: Negative. For each class, 250 clips: Training; 100 clips: Testing

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Dataset: Our dataset consists of 700 movie clips (350 positive and 350 negative) along with dialogs collected from 17 movies belonging to 5 genres viz. Comedy(6), Action(2), Adventure (3), Drama(1), Horror(5). The clips were labeled by 20 volunteers as positive or negative. Out of the 350 clips used for each class, 100 were used for training and 250 for testing.

Experiments using visual cues alone: For FER, only speaker's face is considered. Speaker is identified by lip motion analysis[10]. Experiments were conducted to measure the stand alone efficacy of the FER algorithm without using the lexical cue. To avoid errors due to variation in scale of the faces, features are normalized using the Euclidean distance between the point midway between inner-eye corners and the nose tip. This distance is robust against pose variations in yaw direction, which are found to be more prominent. Figure 2 shows 5 most relevant features selected , marked in red. An SVM classifier with an RBF kernel is trained with the selected features (Section 3). The parameters of the RBF kernel used are $C = 4096$ and $\gamma = 0.125$.

A test clip is preprocessed, relevant features are extracted in the same manner as for training sequences and tested against the classifier. Results of FER experiments, averaged over 10 runs are presented as a confusion matrix in Table 2 (a).

Experiments using lexical cues: Out of 700 clips for 90 instances semantic orientation could not be calculated because dialogues were very short. In such cases, lexical cues were given zero weight in the fusion process. Results for classification of dialogues as having positive or negative emotions is given in Table 2(b). The result is better than the visual cues probably because of the variations of pose, scale, expression intensity, rigid motion etc. in the video sequences.

Experiments combining visual and lexical cues: Results of fusion are shown in Table 2(c). The results shown for all the three cases in Table 2 are averaged over 10 runs. The test and training set are randomly sampled. Note that fusion improves the recognition accuracy.

Fusion depends on the dynamic weightage given to each

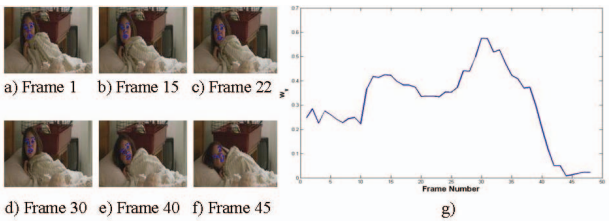

Fig. 3: Variation of visual weight with head pose. Observe the sudden reduction of weight at around frame 40 as the face tilts to the left.

cue. Variation of the weight on visual cues with the progression of a test sequence is given in Figure 3. A sudden decrease of weight is observed with a head tilt, thus reducing the possibility of errors due to variations in visual cues. There are examples also where either one of the two modalities were insufficient for predicting the emotion. However fusing the two modalities gave correct result (See Table 3).

5. DISCUSSIONS

This paper presents a dynamic weighting based multi-modal approach to fusing lexical and visual cues in order to recognize positive and negative emotions of actors in movies. It was observed that for ER, lexical cues performed better (86.7%) as compared to visual cues (76.7%) possibly due to variations in image appearance. An improvement was observed in ER performance by fusing the two cues (ARR = 92.7%) as compared to using any of the two cues independently. The dynamic weights are modeled based on physical and contextual cues. We note that the weights could also be discriminatively learnt from the data but the number of contextual possibilities are too many.

Several issues will be considered for future research. Detecting and localizing FFPs for non-frontal faces is still difficult. In extracting lexical cues, we observe problems due to different meanings of a word and due to different ways of saying the same message. Using the context can be helpful for determining SO under word meaning disambiguation.

6. REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] Z.H. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," vol. 31, no. 1, pp. 39–58, January 2009.

Table 3: An instance showing effectiveness of fusion. Lexical weight is high due to high SO magnitude. Lower weight is assigned to FER result due to small size of the face, low illumination and low facial expression intensity. Upon fusion, the correct prediction of LA dominates the incorrect prediction by FER.

Sample Frame	
FER	Negative
Visual weight	0.21
Text	Captain! Thank goodness you're back.
LA	Positive(SO=+0.32)
Lexical weight	0.79
FER+LA	Positive

[3] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[4] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[5] C.M. Lee and S.S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[6] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 417–424.

[7] A. Hanjalic and L.-Q.Xu., "Affective video content representation and modeling.," *IEEE Transaction on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[8] M. Jones and P. Viola, "Fast multi-view face detection," in *CVPR*, 2003.

[9] T.D. Nguyen and S. Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language," in *Proc. Conf. Face & Gesture Recognition 'FG08'*, 2008, pp. 1–7.

[10] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," in *Proceedings of the British Machine Vision Conference*, 2006.